STORAGE SYSTEM AND CONTROL METHOD THEREFOR

Patent number: JP2003162439

Publication date: 2003-06-06

Inventor: SONODA KOJI; MATSUNAMI NAOTO; KITAMURA MANABU; TAKADA YUTAKA

Applicant: HITACHI LTD

Classification:

- international: ***G06F3/06; G06F7/00; G06F12/00; G06F17/30;*** **G06F3/06; G06F7/00; G06F12/00; G06F17/30;** (IPC1-7): G06F12/00; G06F3/06

- european: G06F3/06M

Application number: JP20010358320 20011122

Priority number(s): JP20010358320 20011122

Also published as:

EP1315074 (A2)
US6850955 (B2)
US2003105767 (A1)

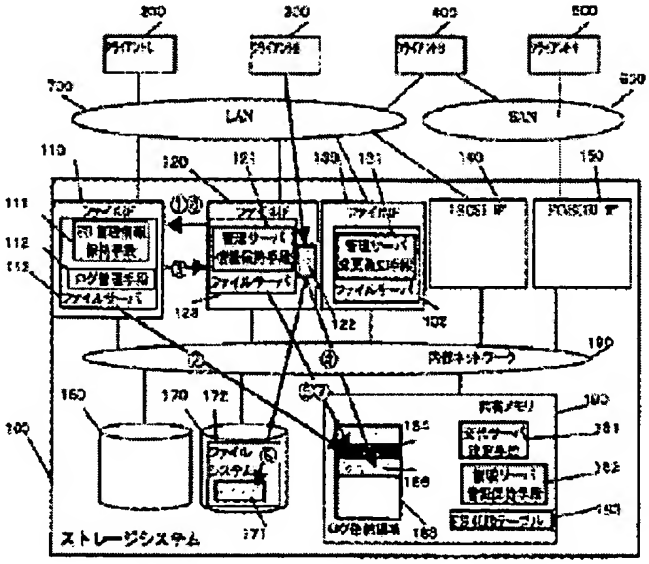
Abstract of **JP2003162439**

PROBLEM TO BE SOLVED: To provide both interfaces of SAN and NAS for elimination of losing data, when troubles occur, and to enable accessing highly efficiently file systems which have arbitrary number of the same NAS interfaces.

SOLUTION: This system is provided with a plurality of interfaces for connecting to the outside, a plurality of disks 160, 170 which can be accessed by the plurality of interfaces, and a common memory 180 which can be accessed by the plurality of interfaces. The plurality of interfaces are block interfaces 140, 150 which process I/O requests on disk block units, and file interfaces mounting file servers 110, 120, 130 which process I/O requests on file units. In a part of the disks, a file system 172 which the file server shares with is constructed; and in the shared memory, a log storing domain which maintains the change logs of the file system and a management file-server information storing domain, which maintains information relating the management file-server which performs exclusive control of the file system and management of the log storing domain are constructed.

COPYRIGHT: (C)2003,JPO

図1



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－162439
（P2003－162439A）
(43)公開日　平成15年６月６日(2003.6.6)

(51)Int.Cl.⁷	識別記号	ＦＩ	テーマコード（参考）
Ｇ０６Ｆ　12/00	５４５	Ｇ０６Ｆ　12/00	５４５Ａ　５Ｂ０６５
	５３１		５３１Ｊ　５Ｂ０８２
	５３５		５３５Ｂ
3/06	３０１	3/06	３０１Ａ
			３０１Ｚ

審査請求　未請求　請求項の数11　ＯＬ　（全 17 頁）　最終頁に続く

(21)出願番号　特願2001－358320(P2001－358320)

(22)出願日　平成13年11月22日(2001.11.22)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　薗田　浩二
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　松並　直人
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

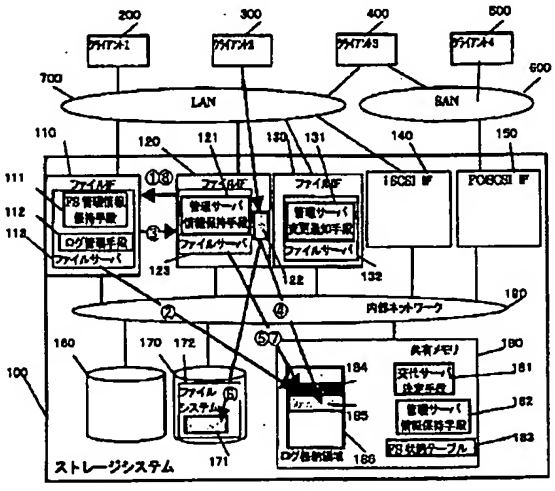
(74)代理人　100093492
弁理士　鈴木　市郎　（外１名）

最終頁に続く

(54)【発明の名称】　ストレージシステム及びその制御方法

(57)【要約】
【課題】　ＳＡＮ、とＮＡＳとの両インタフェースを提供し、障害発生時にもデータが失われることをなくし、また、任意の数のＮＡＳインタフェースが同一のファイルシステムに高性能アクセスすることを可能にする。
【解決手段】　外部接続用の複数のインタフェースと、複数のインタフェースからアクセス可能な複数のディスク１６０、１７０と、複数のインタフェースからアクセス可能な共有メモリ１８０とを備える。複数のインタフェースは、ディスクブロック単位のＩ／Ｏ要求を処理するブロックインターフェース１４０、１５０と、ファイル単位のＩ／Ｏ要求を処理するファイルサーバ１１０、１２０、１３０を搭載したファイルインタフェースとである。ディスクの一部には、ファイルサーバが共有するファイルシステム１７２が構築され、前記共有メモリには、前記ファイルシステムの変更ログを保持するログ格納領域と、前記ファイルシステムの排他制御及びログ格納領域の管理を行う管理用のファイルサーバに関連する情報を保持する管理ファイルサーバ情報格納領域とが構築されている。

図1



【特許請求の範囲】

【請求項１】　外部ネットワーク接続用の複数のインタフェースと、該複数のインタフェースからアクセス可能な複数のディスクと、前記複数のインタフェースからアクセス可能な共有メモリとを備えたストレージシステムにおいて、前記複数のインタフェースには、ディスクブロック単位のＩ／Ｏ要求を処理するブロックインターフェースと、ファイル単位のＩ／Ｏ要求を処理するファイルサーバを搭載したファイルインタフェースとのいずれかあるいは両方が搭載され、前記複数のディスクの一部には、複数のファイルサーバが共有アクセス可能なファイルシステムが構築され、前記共有メモリには、前記ファイルシステムの変更ログを保持するログ格納領域と、前記ファイルシステムの排他制御及びログ格納領域の管理を行う管理用のファイルサーバに関連する情報を保持する管理ファイルサーバ情報格納領域とが構築されていることを特徴とするストレージシステム。

【請求項２】　前記変更ログは、前記ファイルシステムの変更メタデータと、Ｉ／Ｏ要求に含まれる書き込みデータとの両方により構成されることを特徴とする請求項１記載のストレージシステム。

【請求項３】　前記管理ファイルサーバ情報は、前記ログ格納領域に前記変更メタデータのみを格納するか、変更メタデータと書き込み要求データとの両方を格納するかを示す設定情報であることを特徴とする請求項２記載のストレージシステム。

【請求項４】　前記ファイルシステムの管理ファイルサーバ以外のファイルサーバは、外部ネットワークから受信したファイルＩ／Ｏ要求がアクセスするファイルを格納しているファイルシステムの管理を行う管理ファイルサーバに対して、ファイル識別情報とアクセス領域情報とを含むファイルアクセス情報を送信し、その応答としてディスクブロック情報とログ格納アドレス情報とを受信する手段と、前記ファイルＩ／Ｏ要求に含まれる書き込みデータをログ格納アドレス情報に基づいて前記共有メモリに格納する手段と、前記書き込みデータを前記ディスクブロック情報に基づいて前記ファイルシステムが構築されているディスクに格納する手段とを備えることを特徴とする請求項２または３記載のストレージシステム。

【請求項５】　前記管理ファイルサーバは、他のファイルサーバからファイルアクセス情報を受信する手段と、前記受信したファイルアクセス情報を用いて該当するファイルをロックし、ディスクブロックを割り当て、該当するディスクブロック情報を算出するファイルシステムの管理手段と、前記ファイルアクセス情報を用いてログ格納領域内のログ格納アドレスを割り当てるログ格納領域管理手段と、ディスクブロック情報及びログ格納アドレス情報を、ファイルアクセス情報を送信してきたファイルサーバに送信する手段とを備えることを特徴とする請求項４記載のストレージシステム。

【請求項６】　前記管理ファイルサーバは、前記ログ格納領域のサイズを設定するインタフェースを備えることを特徴とする請求項１ないし５のうちいずれか１記載のストレージシステム。

【請求項７】　前記ファイルシステムの管理ファイルサーバ以外のファイルサーバは、前記管理ファイルサーバの障害発生時、前記管理ファイルサーバが管理していたログ格納領域に格納されている変更ログを用いてファイルシステムを復旧する手段を備えることを特徴とする請求項１ないし６のうちいずれか１記載のストレージシステム。

【請求項８】　外部ネットワーク接続用の複数のインタフェースと、該複数のインタフェースからアクセス可能な複数のディスクと、前記複数のインタフェースからアクセス可能な共有メモリとを備えたストレージシステムの制御方法において、

前記複数のインタフェースには、ディスクブロック単位のＩ／Ｏ要求を処理するブロックインターフェースと、ファイル単位のＩ／Ｏ要求を処理するファイルサーバを搭載したファイルインタフェースとのいずれかあるいは両方が搭載され、前記複数のディスクの一部には、複数のファイルサーバが共有アクセス可能なファイルシステムが構築され、前記共有メモリには、前記ファイルシステムの変更ログを保持するログ格納領域と、前記ファイルシステムの排他制御及びログ格納領域の管理を行う管理用のファイルサーバに関連する情報を保持する管理ファイルサーバ情報格納領域とが構築されており、

前記ファイルシステムの管理ファイルサーバ以外のファイルサーバは、外部ネットワークからファイル書き込み要求を受信して、そのファイル書き込み要求を解析して書き込み対象ファイルが含まれるファイルシステムの管理ファイルサーバを特定し、管理ファイルサーバに対してファイル書き込み情報を送信した後、その応答としてユーザデータを書き込むディスクブロック情報とログ格納領域内に割り当てられたログ格納アドレス情報とを受信し、受信したログ格納アドレス情報を用いてユーザデータ格納領域にユーザデータを格納した後、ログ格納領域内のログステータス情報を変更し、ディスクブロック情報に基づいて前記ディスクにユーザデータを格納した後、ログ格納領域内のログステータス情報を変更し、ファイル書き込み結果情報を前記ファイルシステムの管理ファイルサーバに送信した後、外部ネットワークを経由して受信したファイル書き込み要求の応答を、外部ネットワークに送信することを特徴とするストレージシステムの制御方法。

【請求項９】　前記管理ファイルサーバは、管理ファイルサーバ以外のファイルサーバからファイル書き込み情報を受信すると、書き込み対象のファイルをロックし、前記ディスク上にユーザデータを書き込むディスクブロ

ックを割り当て、ログ格納領域内のログ格納アドレスを割り当て、該ログ格納領域にファイルシステム管理データの変更情報とログステータス情報とを格納した後、前記ファイル書き込み情報を送信してきた管理ファイルサーバ以外のファイルサーバに、割り当てたディスクブロック情報とログ格納領域情報とを送信し、前記ファイル書き込み情報を送信してきた管理ファイルサーバ以外のファイルサーバからファイル書き込み結果情報を受信すると、その書き込み対象ファイルのロックを解除することを特徴とする請求項８記載のストレージシステムの制御方法。

【請求項１０】 前記管理ファイルサーバ以外のファイルサーバは、前記管理ファイルサーバに障害が発生したことを認識した場合、前記管理ファイルサーバ情報を参照して、前記管理ファイルサーバが管理していたログ格納領域を特定し、特定したログ格納領域に格納されている変更ログを順次参照し、この変更ログが未反映のファイルシステムに対し変更処理を反映させ、全ての変更ログの反映処理の完了後、前記管理ファイルサーバが管理していた全てのファイルシステムの排他制御とログ管理領域とを引き継ぎ、管理ファイルサーバが変更されたことを前記管理ファイルサーバ以外の他のファイルサーバに通知することを特徴とする請求項８または９記載のストレージシステムの制御方法。

【請求項１１】 前記変更ログの反映処理は、変更処理がファイル書き込みの場合、前記ログステータス情報を参照し、ログステータス情報がユーザデータ未書き込み状態となっていれば、ファイルシステムへの変更反映処理は行わず、前記ログステータス情報がユーザデータディスク書き込み完了状態となっていれば、その変更ログに応じて、ファイルシステム上のファイルシステム管理データのみを変更し、前記ログステータス情報がユーザデータログ書き込み完了状態となっていれば、変更ログに含まれるユーザ書き込みデータをファイルシステムに反映した後、ファイルシステム上のファイルシステム管理データを変更する処理であることを特徴とする請求項１０記載のストレージシステムの制御方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ストレージシステム及びその制御方法に係り、特に、ＮＡＳ機能及びＳＡＮ機能を搭載したRストレージシステム及びその制御方法に関する。

【０００２】

【従来の技術】近年のインターネット技術の進歩に伴い、Ｗｅｂアプリケーション、ストリーミングアプリケーション、Ｅビジネスアプリケーション等が急激に普及し、これらのアプリケーションが必要とするデータ量も飛躍的に増加してきており、日常生活やビジネスが要求するストレージ容量は爆発的に増加している。そして、IBM Redbooksの“Storage Networking Virtualization”によれば、ストレージコストは確実に低下しているが、データの管理コストが増大するという問題点が発生してきている。

【０００３】前述の問題点を解決する従来技術として、Storage Area Network（ＳＡＮ）と呼ばれる技術とNetwork Attached Storage（ＮＡＳ）と呼ばれる技術が知られている。前述した“Storage Networking Virtualization”によると、ＳＡＮは、ストレージ専用の高速ネットワークをFibre Channel を用いて構築し、分散されたデータを集約することにより管理コストを低減することができるようにしたものである。この技術は、ストレージ専用のネットワークを使用することにより、Local Area Network（ＬＡＮ）におけるネットワークトラフィックの影響を排除することができるため、高性能なＩ／Ｏ性能を実現することができる。但し、ＳＡＮが提供する機能は、ディスクブロックレベルの入出力機能であるため、異なるホストやＯＳ相互間でファイルレベルでデータ共有することが困難であるという側面を持つ。

【０００４】一方、http://www.sun.com/Storage/white-papers/NAS.htmlの“NETWORK STORAGE SOLUTIONS White Paper Network-Attached Storage”によれば、ＮＡＳは、ＮＦＳやＣＩＦＳ等のネットワークプロトコルを通じたプラットフォーム独立のストレージ共有手段を提供するファイルサーバである。ＮＡＳは、ＳＡＮと同様に、分散されたデータを集約して管理コストを低減することができる。また、ＮＡＳは、ファイルサーバとして最適化されており、ＬＡＮに直接接続され、ファイルレベルのアクセス手段を提供している。

【０００５】前述したように、ＳＡＮ、ＮＡＳは、共にデータを集約し、管理コストを低減するための技術であるが、提供するアクセス手段が異なるため、用途に応じた使い分けが必要となる。また、前述のように、データを集約し多くのホストで共有利用する場合、高信頼性と高可用性とが重要な要素となる。

【０００６】前述したように、ＳＡＮは、ストレージ専用のネットワークであり、その信頼性は、ＳＡＮに接続される個々のストレージに依存し、ストレージとしてＲＡＩＤ（Redundant Array of Inexpensive Disks）を用いることにより高い信頼性を提供することができる。また、ＲＡＩＤが複数のインタフェースを提供することにより、あるインタフェースに障害が発生した場合にも、別のインタフェースを用いてサービスを継続することができ、高い可用性を提供することができる。

【０００７】一方、ＮＡＳは、ファイルシステムを持つファイルサーバであり、ＮＡＳの信頼性は、ファイルサーバの信頼性そのものとなる。ところが、ファイルサーバは、ＲＡＩＤを用いるだけでは高い信頼性を提供することができない。文献“最前線ＵＮＩＸ（登録商標）のカーネル”、ユーレッシュ・ヴァハリアの９．１２．５

節によれば、ＵＮＩＸ（登録商標）のファイルシステムは、性能向上のためにメモリ上にバッファキャッシュを設け、複数の書き込み処理を纏めて、一括したディスク書き込みを行うようにしている。このため、この技術は、システムクラッシュ時にディスクに書き込まれていないデータが失われてしまう。失われるデータは、ファイルデータそのものと、ファイルシステムの構造を記述しているメタデータの２つに分類することができる。メタデータの変更が失われた場合、ファイルシステムに矛盾が生じ、ファイルシステムを使用できなくなるという問題が生じる。

【０００８】このような問題を解決する方法として、“最前線ＵＮＩＸ（登録商標）のカーネル”、ユーレッシュ・ヴァハリアの１１．７節に述べられているメタデータロギングの技術と、“最前線ＵＮＩＸ（登録商標）のカーネル”、ユーレッシュ・ヴァハリアの１１．５節に述べられているログ構造ファイルシステムの技術が知られている。

【０００９】メタデータロギングは、ディスク上に固定的に設けられた領域に、メタデータの変更ログを常に書き込み、システムクラッシュ時にこのメタデータ変更ログを参照して、ディスクに反映されていないメタデータの変更を反映させることにより、ファイルシステムの矛盾を解決するという方法である。メタデータロギングは、このような方法により、ファイルシステムの矛盾を発生させることをなくすことができるが、ファイルデータが失われる可能性は相変わらず残っている。

【００１０】ログ構造ファイルシステムは、ファイルシステムに対して行われる変更をメモリ上の巨大なログエントリに蓄積し、ファイルシステムの一貫性を保持した状態で常にディスク上のログの最後尾に書き込むことにより、メタデータとユーザデータとの両方を保証することを可能としたものである。但し、このファイルシステムは、ディスク上のログに書き込まれる前にシステムクラッシュが生じた場合、メモリ上のログエントリに保存されていたデータが失われてしまうという問題を生じる。

【００１１】このような問題を解決する技術として、http://www.netapp.com/tech_library/1004.htmlに記述される Network Appliance社の “Using NUMA Interconnectsto Implement Highly Available File Server Appliances”が知られている。この技術は、ＮＡＳ専用に最適化されたもので、不揮発性メモリ（ＮＶＲＡＭ）とＲＡＩＤディスクとを持ち、ＲＡＩＤディスク上にログ構造ファイルシステムを構築したものである。そして、ＮＶＲＡＭには、ネットワーク経由で受信したＮＦＳコマンドを全てロギングし、ＲＡＩＤディスクには、一貫性を保った状態のログを格納することとしている。

【００１２】前述の技術は、ログエントリをＲＡＩＤディスクに書き込む前にシステムがクラッシュした場合にも、システム復旧後にＮＶＲＡＭ上のＮＦＳコマンドログを用いて、ファイルシステム処理を再実行してファイルシステムを完全な状態に復旧することができ、また、データを完全に保証することができる。

【００１３】また、この技術は、ファイルサーバを搭載した２つのノードを独立したネットワークにより接続し、さらに、ＲＡＩＤディスクを両方のノードに接続しておくことにより、片方のノードに障害が発生した場合でも、もう片方が処理を引き継いでサービスを継続して提供するフェイルオーバ機能を提供することができる。さらに、この技術は、各ノードのＮＶＲＡＭに、それぞれ相手ノードのＮＦＳコマンドログのコピーを格納するエリアを確保し、ＮＦＳコマンド受信時に、自ノードのＮＶＲＡＭ上にログを格納すると同時にネットワーク経由で相手ノードのＮＶＲＡＭにもコピーを行っておくため、システム障害発生時に、自ノード上に保存されている相手ノードのＮＦＳコマンドログをアクセスし、相手ノードが使用していたＲＡＩＤディスクのファイルシステムを復旧してサービスを継続することができ、高い可用性を提供することができる。

【００１４】

【発明が解決しようとする課題】前述した従来技術は、ＳＡＮ及びＮＡＳそれぞれの技術により管理コストを低減することができるが、、ＳＡＮへのストレージを提供するＲＡＩＤと、ＮＡＳにおけるファイルサーバとを、それぞれ別の装置として実現しなければならないため、両方の機能が必要な場合、両方の装置を導入する必要があり、別の管理コストを増加させるという問題点があった。

【００１５】また、前述の２ノード構成のＮＡＳにより高信頼性と高可用性とを提供することができる従来技術は、３ノード以上の構成のＮＡＳに適用する場合、各ノードのＮＶＲＡＭにノード数分のログ格納領域を設ける必要があり、この結果、メモリ消費量が大きくなるという問題点を生じさせる。さらに、この従来技術は、ＮＦＳコマンドを受信する度に全ノードのＮＶＲＡＭへのコピーを行う必要があり、ノード数の増加に応じて性能が低下するという問題を生じさせる。さらに、この従来技術は、複数のノードが同一のファイルシステムを変更する場合について配慮されていない。

【００１６】本発明の目的は、前述した従来技術の問題点を解決し、ＳＡＮインタフェースとＮＡＳインタフェースとの両方のを提供し、障害発生時にもデータが失われることのない高信頼性と、かつ、任意の数のＮＡＳインタフェースが同一のファイルシステムに高性能アクセスすることを可能としたＳＡＮ／ＮＡＳを統合したストレージシステム及びその制御方法を提供することにある。

【００１７】また、本発明の他の目的は、同一のファイルシステムにアクセス可能なＮＡＳインタフェースの数

によらず、ファイルシステム変更ログ格納用のメモリサイズを一定にすることができ、かつ、そのサイズをユーザが指定可能なＳＡＮ／ＮＡＳを統合したストレージシステム及びその制御方法を提供することにある。

【００１８】さらに、本発明の他の目的は、あるＮＡＳインタフェースに障害が発生した場合にも、別のＮＡＳインタフェースがその処理を引き継ぐフェイルオーバ処理を可能にし、このフェイルオーバ処理を正常なＮＡＳインタフェースが存在する限り連続して実行することができる高可用性を持ったＳＡＮ／ＮＡＳを統合したストレージシステム及びその制御方法を提供することにある。

【００１９】

【課題を解決するための手段】本発明によれば前記目的は、外部ネットワーク接続用の複数のインタフェースと、該複数のインタフェースからアクセス可能な複数のディスクと、前記複数のインタフェースからアクセス可能な共有メモリとを備えたストレージシステムにおいて、前記複数のインタフェースには、ディスクブロック単位のＩ／Ｏ要求を処理するブロックインターフェースと、ファイル単位のＩ／Ｏ要求を処理するファイルサーバを搭載したファイルインタフェースとのいずれかあるいは両方が搭載され、前記複数のディスクの一部には、複数のファイルサーバが共有アクセス可能なファイルシステムが構築され、前記共有メモリには、前記ファイルシステムの変更ログを保持するログ格納領域と、前記ファイルシステムの排他制御及びログ格納領域の管理を行う管理用のファイルサーバに関連する情報を保持する管理ファイルサーバ情報格納領域とが構築されていることにより達成される。

【００２０】また、前記目的は、外部ネットワーク接続用の複数のインタフェースと、該複数のインタフェースからアクセス可能な複数のディスクと、前記複数のインタフェースからアクセス可能な共有メモリとを備えたストレージシステムの制御方法において、前記複数のインタフェースには、ディスクブロック単位のＩ／Ｏ要求を処理するブロックインターフェースと、ファイル単位のＩ／Ｏ要求を処理するファイルサーバを搭載したファイルインタフェースとのいずれかあるいは両方が搭載され、前記複数のディスクの一部には、複数のファイルサーバが共有アクセス可能なファイルシステムが構築され、前記共有メモリには、前記ファイルシステムの変更ログを保持するログ格納領域と、前記ファイルシステムの排他制御及びログ格納領域の管理を行う管理用のファイルサーバに関連する情報を保持する管理ファイルサーバ情報格納領域とが構築されており、前記ファイルシステムの管理ファイルサーバ以外のファイルサーバが、外部ネットワークからファイル書き込み要求を受信して、そのファイル書き込み要求を解析して書き込み対象ファイルが含まれるファイルシステムの管理ファイルサーバを特定し、管理ファイルサーバに対してファイル書き込み情報を送信した後、その応答としてユーザデータを書き込むディスクブロック情報とログ格納領域内に割り当てられたログ格納アドレス情報とを受信し、受信したログ格納アドレス情報を用いてユーザデータ格納領域にユーザデータを格納した後、ログ格納領域内のログステータス情報を変更し、ディスクブロック情報に基づいて前記ディスクにユーザデータを格納した後、ログ格納領域内のログステータス情報を変更し、ファイル書き込み結果情報を前記ファイルシステムの管理ファイルサーバに送信した後、外部ネットワークを経由して受信したファイル書き込み要求の応答を、外部ネットワークに送信することにより達成される。

【００２１】

【発明の実施の形態】以下、本発明によるストレージシステム及びその制御方法の実施形態を図面により詳細に説明する。

【００２２】図１は本発明の一実施形態によるストレージシステムの構成を示すブロック図、図２はクライアントからのＩ／Ｏ要求を受信したファイルサーバの処理動作を説明するフローチャート、図３は管理ファイルサーバの処理動作を説明するフローチャート、図４は管理ファイルサーバによるファイルシステム復旧の処理動作を説明するフローチャート、図５は管理ファイルサーバ情報保持手段が持つ情報について説明する図である。図１において、１００はストレージシステム、１１０、１２０、１３０はファイルインタフェースボード、１１１はＦＳ（ファイルシステム）管理情報保持手段、１１２はログ格納領域管理手段、１１３、１２３、１３２はファイルサーバ、１２２はデータ一時蓄積部、１４０はｉＳＣＳＩインタフェースボード、１５０はＦＣ／ＳＣＳＩインタフェースボード、１６０、１７０はディスク、１７２はファイルシステム、１８０は共有メモリ、１８１は交代サーバ決定手段、１８２は管理ファイルサーバ情報保持手段、１８３はファイルサーバ状態保持テーブル、１８６はログ格納領域、１９０は内部ネットワーク、２００、３００、４００、５００はクライアントホスト、６００はＳＡＮ、７００はＬＡＮである。

【００２３】本発明の実施形態によるストレージシステム１００は、ＬＡＮ７００及びＳＡＮ６００に接続され、ＬＡＮ７００に接続されるクライアントホスト２００、３００、４００、及び、ＳＡＮ６００に接続されるクライアントホスト４００、５００からネットワークを経由して受信するＩ／Ｏ要求を処理する。このストレージシステム１００は、ファイルサーバ１１３、１２３、１３２を搭載したファイルインタフェースボード１１０、１２０、１３０と、ｉＳＣＳＩコマンドを処理するｉＳＣＳＩインタフェースボード１４０と、ファイバーチャネルを介して受信したＳＣＳＩコマンドを処理するＦＣ／ＳＣＳＩインタフェースボード１５０と、ディス

ク160、170及び共有メモリ180とを備え、前述の各ボードとディスク160、170及び共有メモリ180とが内部ネットワーク190で結合されて構成されている。

【0024】ストレージシステム100は、前述したインタフェースボードをそれぞれ異なるスロットに装着している。なお、各スロットは、任意のインタフェースボードをシステム稼動時に挿抜可能である。従って、本発明の実施形態によるストレージシステム100は、ユーザの用途に応じて各インタフェースボードの割合を動的に変更することができる。

【0025】ディスク160、170は、前述の各インタフェースボードのいずれからもアクセス可能である。そして、ディスク160は、iSCSIインタフェースボード140及びFC/SCSIインタフェースボード150がブロック単位のアクセスを行うためのディスクであり、ディスク170には、ファイルサーバがアクセスするためのファイルシステム172が構築されている。なお、ここでいうディスクは、RAIDを含む論理的なディスクを意味しており、簡単のためにそれぞれ1つずつ記載しているが、実際には任意の数のディスクを搭載することが可能である。

【0026】ファイルサーバ113は、ファイルシステム172の管理ファイルサーバであり、他のファイルサーバがファイルシステム172をアクセスする場合に、必ずファイルサーバ113と通信を行う。ファイルサーバ113は、ファイルシステム管理情報保持手段111とログ管理手段112とを有している。ファイルシステム管理情報保持手段111は、ファイルの属性やファイルが保持するディスクブロックのリスト、ファイルシステム内の未使用ディスクブロックリスト等のメタデータや、ファイルデータを保持しており、ファイルシステムへの変更が発生するたびに内部の情報を更新する。但し、これらのデータは、ディスク170上で複数のディスクブロックにまたがって格納されるため、変更の度にディスクへ格納すると性能が大きく低下する。このため、ディスクへのデータの書き戻しは、クライアントホストからの書き戻し要求の発生時か、または、ある一定時間毎に実行する。

【0027】変更データのディスクへの反映をファイルシステム変更処理と非同期に行うと、ファイルサーバに障害が発生した場合に、ディスクに未反映となっている変更データが失われる危険性がある。本発明の実施形態は、この問題を防止するため、ファイルシステムの変更データのログを、不揮発性の共有メモリ180上に設けたログ格納領域186に格納する。ログとして格納するログエントリは、変更メタデータ184と変更ファイルデータ185とにより構成される。ファイルサーバ113は、ログ管理手段112を用いてログ格納領域186を環状に使用し、ファイルシステムの変更データをディスクに反映した場合に、対応する変更データを含むログエントリを解放する。

【0028】本発明の実施形態によるストレージシステムは、前述のように、ファイルシステムの変更と同期的にログを格納することとしているので、全ての変更情報を残すことができ、また、共有メモリ180が不揮発性であるため、変更データが失われることがない。また、ファイルサーバ113に障害が発生した場合にも、再起動後にログ格納領域186内のログデータを参照しファイルシステムを最新の状態に修復することが可能となる。

【0029】ファイルサーバ123は、ファイルシステム172の管理ファイルサーバではないが、クライアントホストからファイルシステム172上のファイルに対するI/O要求を受信した場合に、ファイルシステム172をアクセスする。ファイルサーバ123は、ファイルシステム172のメタデータ情報を持たないため、ファイルシステム172の管理ファイルサーバであるファイルサーバ113を特定し、必要な処理を依頼する必要がある。I/O要求がメタデータアクセスのみの場合、全てのファイルシステム処理は、ファイルサーバ113で完結する。

【0030】一方、READやWRITEといったファイルデータへのアクセスが必要な処理の場合、メタデータの処理は、ファイルサーバ113で行われ、ファイルデータの転送処理は、ファイルサーバ123で行われる。これにより、管理ファイルサーバの負荷を低減すると共に、ファイルデータのファイルサーバ間コピーを不要とすることができる。

【0031】図5に示すように、管理サーバ情報保持手段182は、ログ格納領域管理テーブル5000とディスク管理テーブル5100とにより構成される。ログ格納領域管理テーブル5000は、管理ファイルサーバの番号5001と、そのログ格納領域アドレス5002と、ログ格納領域サイズ5003とからなる。管理ファイルサーバ113は、ファイルシステム172を正常にマウントするとき、ログ格納領域を割り当て、その先頭アドレス及びサイズをこのテーブル5000に格納し、異常終了後にファイルシステムをマウントするとき、このテーブル5000を参照してファイルシステムの復旧を行う。なお、複数のファイルシステムが備えられる場合に、複数のファイルシステムのそれぞれに対するファイル管理サーバは、異なったものであってもよい。

【0032】ディスク管理テーブル5100は、各ディスクのディスク番号5001、デフォルト管理ファイルサーバ5102、カレント管理ファイルサーバ5103、ログ種別5103を保持している。ストレージシステム100は、管理ファイルサーバのフェイルオーバ機能を提供しており、デフォルト管理ファイルサーバ5102は、フェイルオーバが発生していない場合にそのデ

ィスクを管理すべき管理ファイルサーバが登録される。カレント管理ファイルサーバ５１０３は、フェイルオーバの有無にかかわらず現在そのディスクを管理している管理ファイルサーバが登録される。

【００３３】ログ種別５１０４は、ログに格納する変更データの種類を記述する。変更データの種類としては、変更メタデータ、ファイルデータの２種類があり、ログ種別５１０４にはその両方を格納するか、変更メタデータのみを格納するか、あるいは、全くログを採取しないかの３つのうちの１つが登録される。変更メタデータとファイルデータとの両方をログに格納すれば、全てのデータを完全に保証することができる。変更メタデータのみをログに格納する場合、ファイルシステムの整合性を保証することができるが、ファイルデータが失われる可能性がある。また、全くログを採取しない場合、ファイルシステムの整合性もファイルデータも失われる可能性がある。この場合システム障害発生後、最初にファイルシステムを使用する前に、ファイルシステムの整合性チェックの処理と修復を行うｆｓｃｋのような処理とを行う必要がある。一般に、このｆｓｃｋの処理時間は、ファイルシステムサイズに比例して増加し、数分から数十分を要することがあるため、フェイルオーバ時には使用できない。データの信頼性保証と性能とはトレードオフの関係にあるため、信頼性を向上させるとファイルシステムアクセス性能は低下する。このため、本発明の実施形態では、ログ種別５１０４は、ユーザが用途に応じて指定することができるようにしている。

【００３４】次に、図２に示すフローを参照して、ファイルサーバ１２３がクライアントホスト３００からファイルシステム１７２に対するＩ／Ｏ要求を受信した場合の処理動作を説明する。

【００３５】（１）まず、ファイルサーバ１２３は、クライアントホスト３００から受信したＩ／Ｏ要求を解析し、ファイルシステム１７２へのＩ／Ｏ要求であることを検出すると、ディスク管理テーブル５１００を検索し、ディスク番号５１０１がファイルシステム１７２のディスク番号と一致するエントリを特定し、そのカレント管理ファイルサーバ５１０３に格納されている管理ファイルサーバを読み出す。ここでは、管理ファイルサーバとして、ファイルサーバ１１３が設定されているものとする（ステップ２００１）。

【００３６】（２）次に、管理ファイルサーバ１１３にファイルアクセス要求を送信する。このとき、Ｉ／Ｏ要求がＷＲＩＴＥ要求の場合、要求に伴って受信しているファイルデータは、ファイルサーバ１２３内のデーター時蓄積部１２２に残し、ファイルサーバ１１３には送信しない（ステップ２００２（図１の❶））。

【００３７】（３）次に、Ｉ／Ｏ要求がＲＥＡＤやＷＲＩＴＥ等のファイルデータをアクセスする要求か否かを判断する。ファイルデータをアクセスしないＩ／Ｏ要求の場合、その処理は、管理ファイルサーバ１１３のみで完結するので、その場合、ファイルサーバ１２３は、ファイルサーバ１１３からの処理結果を受信した後、その処理結果をクライアントホスト３００に返信して処理を終了する（ステップ２００３、２０１４、２０１５）。

【００３８】（４）ステップ２００３の判定で、Ｉ／Ｏ要求がファイルデータをアクセスする要求の場合、ファイルサーバ１１３からファイルデータを保存するディスクブロック番号と、ログ格納アドレス情報とを受信する。このとき、ログ格納アドレス情報の受信は、ＷＲＩＥ要求の場合のみであり、ＲＥＡＤ要求のときには不要である（ステップ２００４（図１の❸））。

【００３９】（５）次に、Ｉ／Ｏ要求がＷＲＩＴＥ要求であるか否かを判断する。Ｉ／Ｏ要求がＷＲＩＴＥでな居場合、すなわち、ＲＥＡＤの場合、ファイルデータの格納されているディスクブロック番号のみをファイル管理サーバ１１３から受信しているため、受信したディスクブロック番号を使用してディスク１７０をアクセスしてファイルデータを読み出し、読み出したファイルデータをクライアントホスト３００に返信し、その後、処理結果をファイルサーバ１１３に返信して処理を終了する（ステップ２００５、２０１２、２０１３、２０１１）。

【００４０】（６）ステップ２００５の判定で、Ｉ／Ｏ要求がＷＲＩＴＥであった場合、ファイルサーバ１１３から受信したログ格納アドレス情報を用い、ファイルデータをデーター時蓄積部１２２から変更ファイルデータ格納領域１８５に格納する。その後、変更メタデータ格納領域１８４に格納されているログステータス情報を“データログ書き込み”の状態に変更する（ステップ２００６（図１の❹）、２００７（図１の❺））。

【００４１】（７）ここまでの処理により、クライアント３００からのＩ／Ｏ要求の全てがログ格納領域１８６に反映されたため、以後、ファイルサーバ１１３に障害が発生しても、ログ情報を用いてファイルシステム１７２を復旧することができる。このため、ファイルサーバ１２３は、クライアントホスト３００に対して処理結果を返信する（ステップ２００８）。

【００４２】（８）次に、ファイルサーバ１２３は、ファイルサーバ１１３から受信したディスクブロック情報を用いて、データー時蓄積部１２２に格納していたファイルデータをディスク１７０に格納し、その後、ログステータス情報を“データディスク書き込み”状態に変更し、さらに、ファイルサーバ１１３に処理結果を返信して処理を終了する（ステップ２００９（図１の❻）～２０１１（図１の❽））。

【００４３】ファイルサーバ１２３によるＩ／Ｏ要求に対する処理は、前述した通りであるが、ファイルサーバ１２３以外にファイルシステム１７２をアクセスするファイルサーバが存在した場合にも、ファイルデータのロ

グ格納領域１８６への格納及びディスク１７０へのアクセス処理は、それぞれ並列に実行することができるため高速なＩ／Ｏ処理性能を提供できる。

【００４４】次に、図３に示すフローを参照して、管理ファイルサーバ１１３での処理動作を説明する。ここでは、ファイルサーバ１２３からのファイルアクセス要求を受信した場合の処理について説明するが、その他のファイルサーバからファイルアクセス要求を受信した場合も同様である。

【００４５】（１）管理ファイルサーバ１１３は、ファイルサーバ１２３からのファイルアクセス要求を受信すると、そのファイルアクセス要求を解析し、ファイルシステム管理情報保持手段１１１を用いてアクセスするファイルを特定すると同時に、ファイルに対してロックをかける（ステップ３００１）。

【００４６】（２）次に、ファイルアクセス要求がＲＥＡＤやＷＲＩＴＥ等のファイルデータアクセス要求か、それ以外の要求かを判断する。要求がファイルデータアクセス要求以外の場合、ファイルサーバ１２３にディスクブロック番号やログ格納アドレス情報を返信する必要がなく、自サーバによるメタデータアクセスのみで処理が終了する。このため、管理ファイルサーバ１１３は、メタデータアクセスを行い、メタデータの変更があった場合に、ログ管理手段１１２が割り当てたログエントリに変更メタデータを格納する。その後、ファイルのロックを解除し、処理結果をファイルサーバ１２３に返信して処理を終了する（ステップ３００２、３００９～３０１１）。

【００４７】（３）ステップ３００２の判定で、要求がファイルデータアクセス要求であった場合、次に、Ｉ／Ｏ要求がＷＲＩＴＥ要求であるか否かを判断する。ＷＲＩＴＥ要求でなかった場合、すなわち、ＲＥＡＤ要求であった場合、ファイルシステムに対する変更が発生しないためログ採取の必要がない。従って、管理ファイルサーバ１１３は、メタデータをアクセスし、ファイルデータが格納されているディスクブロック番号を算出し、ファイルサーバ１２３にディスクブロック番号を返信する（ステップ３００３、３０１２、３０１３）。

【００４８】（４）ステップ３００３の判定で、Ｉ／Ｏ要求がＷＲＩＴＥ要求であった場合、ファイルデータを格納するディスクブロックを割り当てると共に、メタデータの変更を行う。次に、変更メタデータとＷＲＩＴＥするファイルデータを格納するログエントリサイズを算出し、ログ管理手段１１２を用いて、ログ格納領域１８６内に変更メタデータ格納領域１８４とファイルデータ格納領域１８５とを含むログエントリを割り当てる（ステップ３００４、３００５）。

【００４９】（５）その後、変更メタデータ内のログステータス情報を“データ未書き込み”に設定した後、変更メタデータを変更メタデータ格納領域１８４に格納し、次に、変更メタデータ格納領域１８４及びファイルデータ格納領域１８５のアドレス及びサイズと、ステップ３００４の処理で割り当てたディスクブロック番号をファイルサーバ１２３に返信する（ステップ３００６（図１の❷）、３００７（図１の❸））。

【００５０】（６）ステップ３００７またはステップ３０１３の処理後、ファイルサーバ１２３からのファイルアクセス終了のメッセージを受信すると、ファイルのロックを解除して処理を終了する（ステップ３００８）。

【００５１】次に、図４に示すフローを参照して、管理ファイルサーバ１１３がファイルシステム１７２を復旧する処理動作を説明する。ここでは、管理ファイルサーバ１１３に障害が発生し、その後、管理ファイルサーバ１１３が障害から回復し、ファイルシステム１７２を復旧するものとしている。

【００５２】（１）管理ファイルサーバ１１３は、まず、ファイルシステム１７２を格納されていいるディスク番号を用いてディスク管理テーブル５１００を検索し、ディスク１７０を格納しているエントリを特定し、そのカレント管理ファイルサーバフィールド５１０３を参照して、直前にファイルシステム１７２を管理していた管理ファイルサーバを特定する。ここでは、ファイルサーバ１１３が格納されているものとする。さらに、ログ種別フィールド５１０４を参照して、ログ種別を取得する（ステップ４００１）。

【００５３】（２）次に、ステップ４００２で取得したログ種別が“ログ不採取”であるか否かを判断し、“ログ不採取”の場合、ログを用いた復旧が不可能であるため、ファイルシステムチェックプログラムであるｆｓｃｋを起動し、ファイルシステムの修復を行った後、ファイルシステム復旧処理を終了する（ステップ４００２、４０１０）。

【００５４】（３）ステップ４００２の判定で、ログ種別が“ログ不採取”でなかった場合、ログ格納領域管理テーブル５０００を検索し、管理ファイルサーバ１１３のログ格納領域のアドレス及びサイズを求める。その後、ログ格納領域１８６をスキャンして、ログエントリのポインタをディスク１７０に未反映の変更ログを保持しているログエントリに設定する（ステップ４００３、４００４）。

【００５５】（４）次に、ファイルシステム１７２に対する全ての変更ログをディスク１７０に反映したか否かをチェックし、全て反映済みであれば、ファイルシステム復旧処理を終了する。また、まだ未反映のログエントリが存在する場合、ポインタがさしているログエントリを参照し、そのログステータス情報が“データログ書き込み”状態か否かを調べる（ステップ４００５、４００６）。

【００５６】（５）ステップ４００６の判定で、ログステータス情報が“データログ書き込み”状態であった場

合、ファイルデータがログに格納されているがディスクには未反映の状態を示しているため、ログに格納されているファイルデータを、メタデータ変更ログのディスクブロック情報に従ってディスクに格納する（ステップ４００７）。

【００５７】（６）ステップ４００７の処理後、変更メタデータをファイルシステムの管理自に反映させ、ファイルシステム１７２に関する次の未反映ログエントリにポインタを進める。そして、ステップ４００５の処理に戻って、ここからの処理を繰り返す（ステップ４００８、４００９）。

【００５８】（７）また、ステップ４００６の判定で、ログステータス情報が“データログ書き込み”状態でなかった場合、ファイルデータ格納ログ領域内のデータは意味が無いため、ステップ４００７の処理を実行することなく、ステップ４００８からの処理を実行する。

【００５９】図４により説明した処理を、全てのログエントリについて繰り返して実行することにより、ファイルシステムを最新の状態に復旧することができる。ログ格納領域１８６に格納されている全てのログエントリの処理が終了すると、ログ格納領域１８６内の全てのログエントリは解放される。

【００６０】前述したような構成を持ち、前述したよあな処理を行う本発明の実施形態によれば、ＳＡＮインタフェースとＮＡＳインタフェースとの両方を任意の割合で提供し、障害発生時にもデータを失わない高信頼性を得ることができ、かつ、任意の数のＮＡＳインタフェースが同一のファイルシステムに高性能なアクセスをすることを可能としたＳＡＮ／ＮＡＳ統合ストレージシステムを提供することができる。

【００６１】図６はログ格納領域のサイズをクライアントホストから設定する際の管理ファイルサーバの処理動作の例を説明するフローチャートであり、次に、これについて説明する。

【００６２】（１）管理ファイルサーバ１１３は、クライアントホストからのログ格納領域サイズ設定コマンドを受信すると、受信したサイズの値が正常範囲内であるか否かかを判定し、サイズが正常範囲外であれば、処理結果としてエラーをクライアントホストに返信して、ここでの処理を終了する（ステップ６００１、６００２、６００８）。

【００６３】（２）ステップ６００２の判定で、ログ格納領域サイズが正常範囲内であることが確認できると、ログ格納領域１８６が使用中（領域内にデータがある）か否かをログ管理手段１１２により調べて判定する。ログ格納領域が使用中の場合、次に、新たなＩ／Ｏ要求処理の開始を抑止すると共に、現在実行中のＩ／Ｏ処理終了を待ち合わせる（ステップ６００３、６００４）。

【００６４】（３）その後、ステップ６００５でファイルシステム管理情報保持手段１１１内にあるディスク未反映データを全てディスクに反映させる。その結果としてログ格納領域１８６内の全てのログエントリは解放され、ログ格納領域１８６が未使用の状態と等しくなる（ステップ６００５）。

【００６５】（４）ステップ６００５の処理のご、または、ステップ６００３での判定で、ログ格納領域１８６が使用中でなかった場合、指定されたサイズのログ格納領域を共有メモリ１８０内に確保し、そのアドレスとサイズとをログ格納領域管理テーブル５０００内のログ格納領域アドレスフィールド５００１とログ格納領域サイズフィールド５００２とに格納する。その後、Ｉ／Ｏ処理を再開し、処理結果をクライアントに返信して処理を終了する（ステップ６００６～６００８）。

【００６６】前述した図６に示す例は、ログ格納領域サイズ変更のコマンドを送信したクライアントホストが適切な権限をもっているか否かのチェックを行っていないが、コマンド受信後に適切な権限チェックを行うようにすることもできる。

【００６７】本発明の実施形態は、前述した処理を行うことにより、同一のファイルシステムにアクセス可能なＮＡＳインタフェースの数によらず、ファイルシステム変更ログ格納用のメモリサイズを一定にすることができ、かつ、そのサイズをクライアントが指定した値に設定することが可能となる。

【００６８】図７はファイルサーバ状態テーブル１８３の構成を説明する図、図８はファイルサーバ１３２によるファイルサーバ１１３の監視及びフェイルオーバの処理動作を説明するフローチャートであり、次に、管理ファイルサーバに障害が発生した場合に他のファイルサーバが管理ファイルサーバの処理を引き継ぐフェイルオーバ処理について説明する。

【００６９】ここで説明する例は、図１において、共有メモリ１８０内にある交替サーバ決定手段１８１が、ある管理ファイルサーバの状態を監視し、異常発生時にはその処理を引き継ぐべき交代ファイルサーバを決定し、交代サーバ決定手段１８１によって決定された交代ファイルサーバが、ファイルサーバ状態テーブル１８３を用いて、監視対象の管理ファイルサーバの状態を常に監視し、異常を検知すると即座にフェイルオーバ処理を開始する処理である。説明する例では、管理ファイルサーバ１１３の交代ファイルサーバがファイルサーバ１３２であるとする。

【００７０】ファイルサーバ状態テーブル１８３は、図７に示すように、ファイルサーバ番号７００１、状態７００２、タイムスタンプ７００３、ＩＰアドレスやＭＡＣアドレスなどのネットワーク情報７００４から構成される。各ファイルサーバは、一定のリフレッシュ時間間隔で、ファイルサーバ状態テーブル１８３のタイムスタンプ７００３を更新する。交代ファイルサーバは、監視対象のファイルサーバのタイムスタンプ７００３を一定

のチェック時間間隔で検査し、タイムスタンプ７００３の値が正しく更新されていれば、監視対象のファイルサーバが正しく動作していると判断し、タイムスタンプ７００３の値が更新されていなければ異常が発生したと判断してフェイルオーバ処理を開始する。ここで、チェック時間間隔はリフレッシュ時間間隔よりも大きな値である必要がある。

【００７１】次に、図８に示すフローを参照して、ファイルサーバ１３２によるファイルサーバ１１３の監視及びフェイルオーバ処理の動作を説明する。

【００７２】（１）ファイルサーバ１３２は、チェック時間になるとファイルサーバ状態テーブル１８３から監視対象であるファイルサーバ１１３のエントリを検索して、そのタイムスタンプ７００３をチェックし、タイムスタンプが正しく更新されているか否かを判定する。タイムスタンプが正しく更新されている場合、何も行わずに処理を終了する（ステップ８００１、８００２）。

【００７３】（２）ステップ８００２の判定で、タイムスタンプが正しく更新されていないことを検知すると、ファイルサーバ１３２は、フェイルオーバ処理を開始する。そして、まず、ファイルサーバ状態テーブル１８３のファイルサーバ１１３の状態７００１を“フェイルオーバ処理中”に設定して２重フェイルオーバ処理の起動を抑止する（ステップ８００３）。

【００７４】（３）次に、ディスク管理テーブル５１００を検索し、デフォルト管理ファイルサーバ５１０２がファイルサーバ１１３であるディスクのエントリを取得し、取得したディスクの復旧処理を行う。ここでの復旧処理は、図４により説明した復旧処理と同様に行われる（ステップ８００４、８００５）。

【００７５】（４）ディスクの復旧処理が終了すると、次に、ディスク管理テーブルのカレントファイルサーバに交代ファイルサーバであるファイルサーバ１３２の番号を格納し、ファイルサーバ１１３が管理していた全ディスクの復旧が終了したか否かを判断する。まだ、復旧処理の終了していないディスクが存在する場合、ステップ８００４の処理に戻って以後の処理を繰り返す（ステップ８００６、８００７）。

【００７６】（５）ファイルサーバ１１３が管理していた全てのディスクに関する復旧処理が終了すると、次に、ファイルサーバ状態テーブルのネットワーク情報７００４を参照し、ファイルサーバ１１３のネットワークアダプタに設定されていた情報をファイルサーバ１３２のネットワークアダプタに引き継ぐ（ステップ８００８）。

【００７７】（６）その後、ファイルサーバ状態テーブルにあるファイルサーバ１１３の状態７００２を“フェイルオーバ”状態に変更し、最後に、ファイルサーバ１３２がファイル管理サーバ変更通知手段１３１を用いて他のファイルサーバに管理ファイルサーバの変更を通知してフェイルオーバ処理を終了する（ステップ８００９、８０１０）。

【００７８】前述の処理において、管理ファイルサーバ変更通知を受け取ったファイルサーバ１２３は、管理ファイルサーバ情報保持手段１８２の情報を再び共有メモリ１８０から読み込み、ファイルサーバ１２３内の管理サーバ情報保持手段１２１に格納する。また、交代サーバ決定手段１８１は、交代ファイルサーバ１３２がファイルサーバ１１３の異常を検知した時点で、全ファイルサーバの交代ファイルサーバを再決定し、各ファイルサーバに通知する。この処理により、常に正常な状態のファイルサーバ同士がお互いに監視し合うことが可能となる。

【００７９】本発明の実施形態は、図８により説明した処理を行うことにより、あるＮＡＳインタフェースに障害が発生しても、他に正常なＮＡＳインタフェースが存在する限り、フェイルオーバ処理を用いてサービスを継続でき、高可用性を図ることができる。

【００８０】

【発明の効果】以上説明したように本発明によれば、複数のＮＡＳインタフェースが同一のファイルシステムをアクセスすることができるため、インタフェース数に比例した性能を提供しながら、障害発生時にもデータが失われることのない高い信頼性を得ることができ、さらに、正常なＮＡＳインタフェースが１つでも存在する限り継続してファイルアクセスサービスを行うことができる。

【図面の簡単な説明】

【図１】本発明の一実施形態によるストレージシステムの構成を示すブロック図である。

【図２】クライアントからのＩ／Ｏ要求を受信したファイルサーバの処理動作を説明するフローチャートである。

【図３】管理ファイルサーバの処理動作を説明するフローチャートである。

【図４】管理ファイルサーバによるファイルシステム復旧の処理動作を説明するフローチャートである。

【図５】管理ファイルサーバ情報保持手段が持つ情報について説明する図である。

【図６】ログ格納領域のサイズをクライアントホストから設定する際の管理ファイルサーバの処理動作の例を説明するフローチャートである。

【図７】ファイルサーバ状態テーブルの構成を説明する図である。

【図８】ファイルサーバによる管理ファイルサーバの監視及びフェイルオーバの処理動作を説明するフローチャートである。

【符号の説明】

１００　ストレージシステム

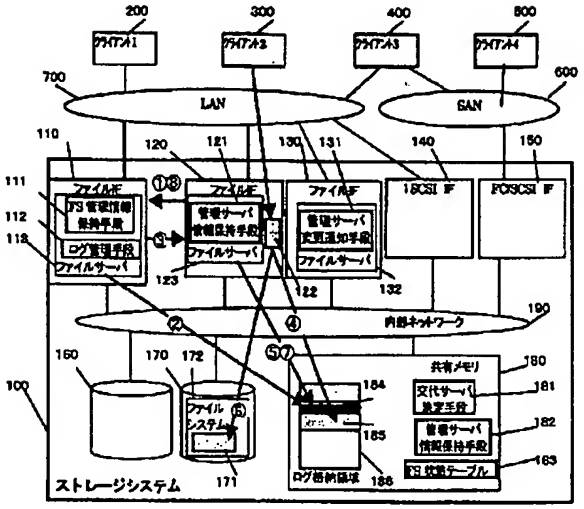
１１０、１２０、１３０　ファイルインタフェースボー

ド

111　FS（ファイルシステム）管理情報保持手段
112　ログ格納領域管理手段
113、123、132　ファイルサーバ
122　データ一時蓄積部
140　iSCSIインタフェースボード
150　FC／SCSIインタフェースボード
160、170　ディスク
172　ファイルシステム
180　共有メモリ
181　交代サーバ決定手段
182　管理ファイルサーバ情報保持手段
183　ファイルサーバ状態保持テーブル
186　ログ格納領域
190　内部ネットワーク
200、300、400、500　クライアントホスト
600　SAN
700　LAN

【図1】

図1



【図7】

図7

183

ファイルサーバ状態テーブル

7001 管理ファイルサーバ番号	7002 状態	7003 タイムスタンプ	7004 ネットワーク情報

【図5】

図5

182

5000

ログ格納領域管理テーブル

5001 管理ファイルサーバ番号	5002 ログ格納領域アドレス	5003 ログ格納領域サイズ

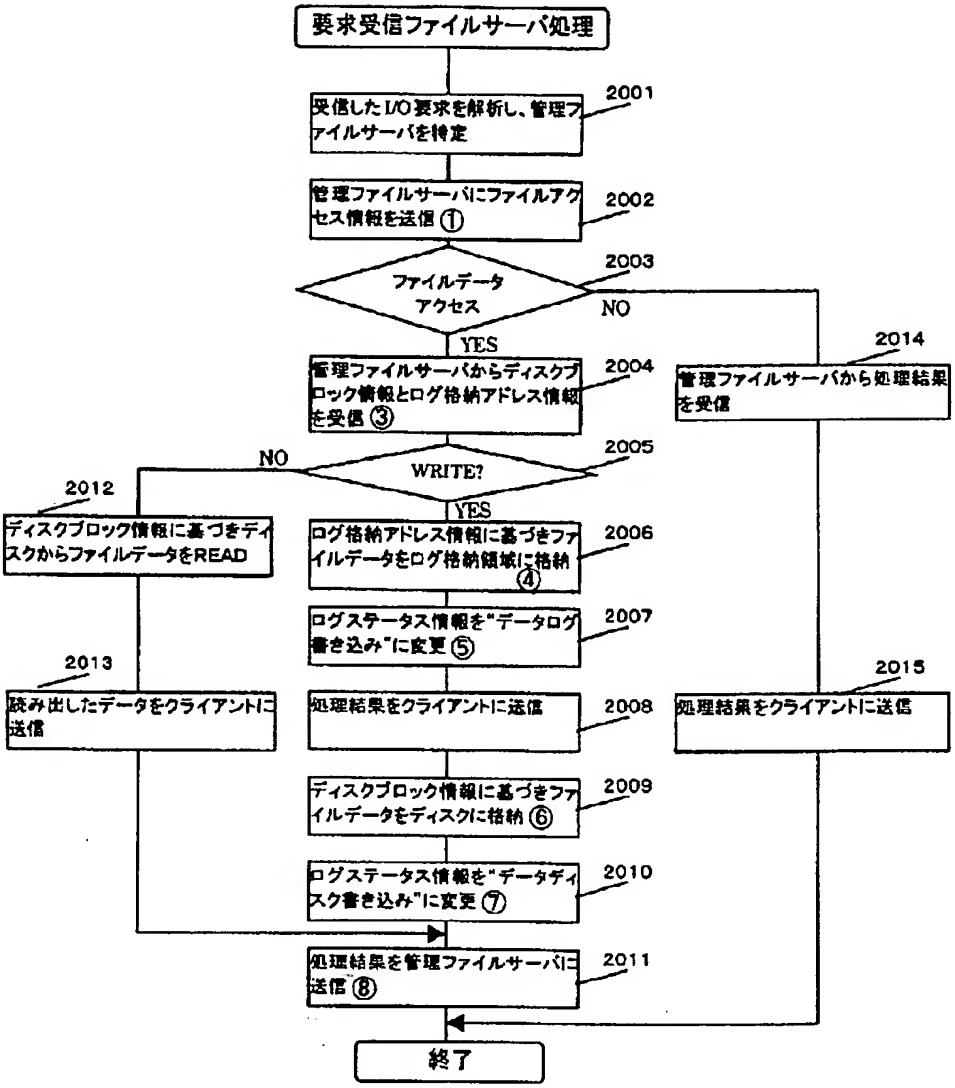
5100

ディスク管理テーブル

5101 ディスク番号	5102 デフォルト管理ファイルサーバ	5103 カレント管理ファイルサーバ	5104 ログ種別

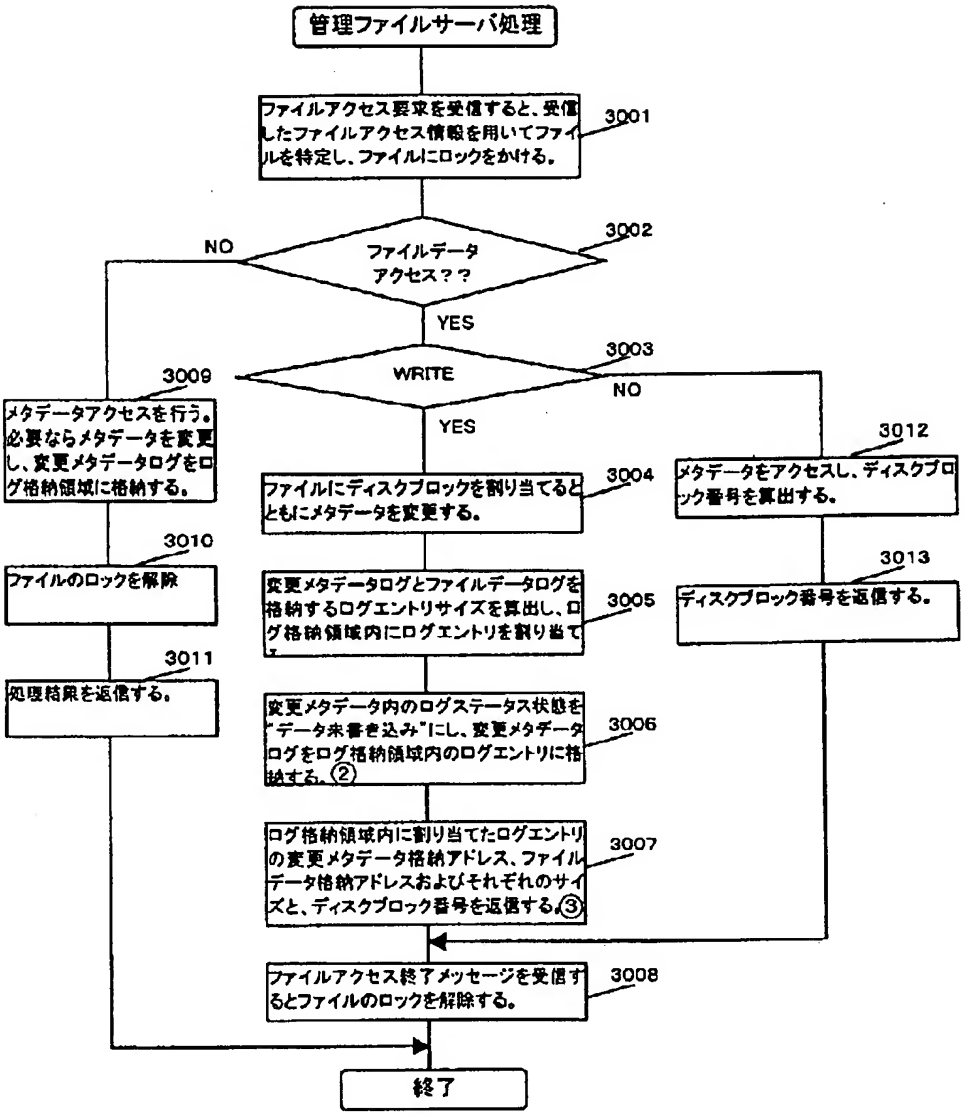
【図2】

図2


要求受信ファイルサーバ処理
受信したI/O要求を解析し、管理ファイルサーバを特定
2001
管理ファイルサーバにファイルアクセス情報を送信①
2002
ファイルデータアクセス
2003
NO
YES
管理ファイルサーバからディスクブロック情報とログ格納アドレス情報を受信③
2004
管理ファイルサーバから処理結果を受信
2014
WRITE?
2005
NO
YES
2012
ディスクブロック情報に基づきディスクからファイルデータをREAD
ログ格納アドレス情報に基づきファイルデータをログ格納領域に格納④
2006
ログステータス情報を"データログ書き込み"に変更⑤
2007
2013
読み出したデータをクライアントに送信
処理結果をクライアントに送信
2008
2015
処理結果をクライアントに送信
ディスクブロック情報に基づきファイルデータをディスクに格納⑥
2009
ログステータス情報を"データディスク書き込み"に変更⑦
2010
処理結果を管理ファイルサーバに送信⑧
2011
終了

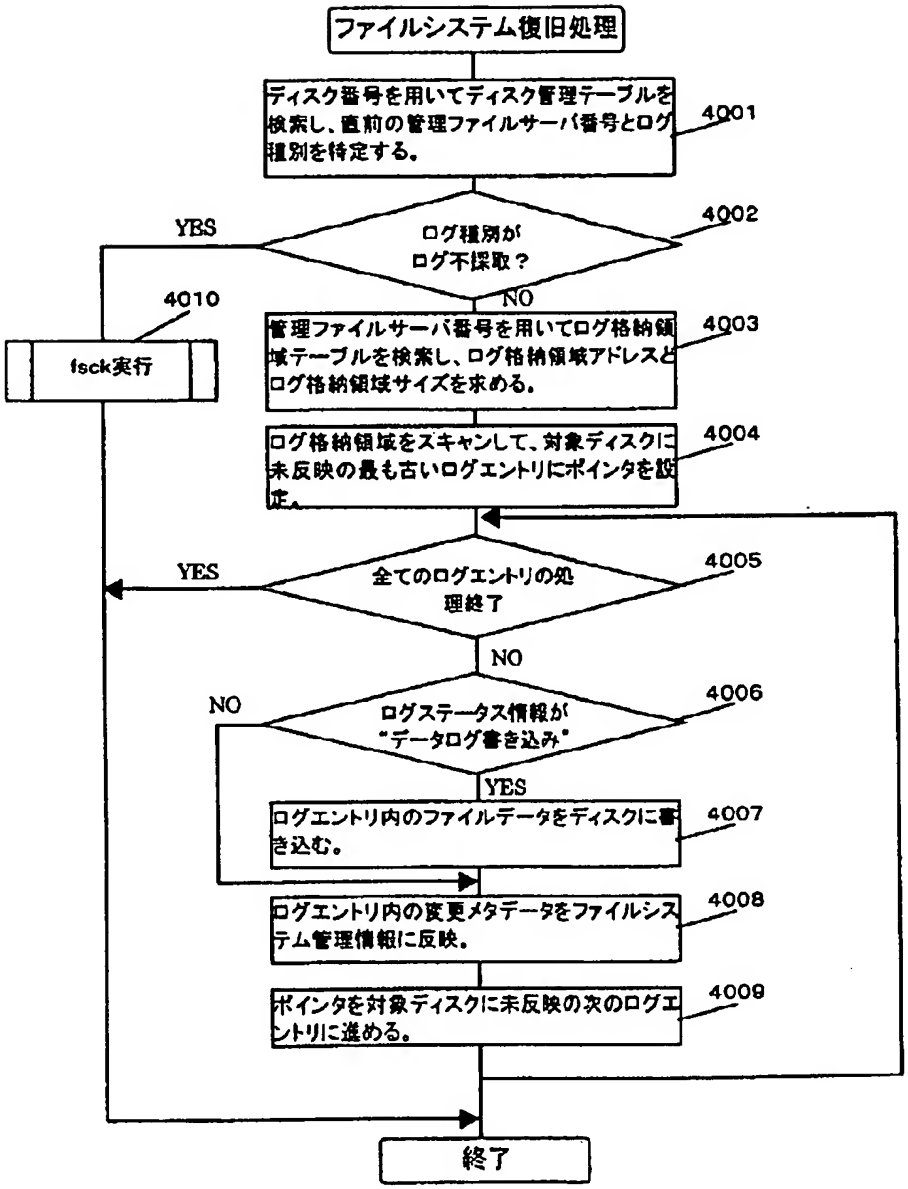
【図3】

図3


管理ファイルサーバ処理
ファイルアクセス要求を受信すると、受信したファイルアクセス情報を用いてファイルを特定し、ファイルにロックをかける。
3001
ファイルデータアクセス？？
3002
NO
YES
WRITE
3003
NO
YES
3009
メタデータアクセスを行う。必要ならメタデータを変更し、変更メタデータログをログ格納領域に格納する。
3010
ファイルのロックを解除
3011
処理結果を返信する。
ファイルにディスクブロックを割り当てるとともにメタデータを変更する。
3004
変更メタデータログとファイルデータログを格納するログエントリサイズを算出し、ログ格納領域内にログエントリを割り当て
3005
変更メタデータ内のログステータス状態を"データ未書き込み"にし、変更メタデータログをログ格納領域内のログエントリに格納する。②
3006
ログ格納領域内に割り当てたログエントリの変更メタデータ格納アドレス、ファイルデータ格納アドレスおよびそれぞれのサイズと、ディスクブロック番号を返信する。③
3007
3012
メタデータをアクセスし、ディスクブロック番号を算出する。
3013
ディスクブロック番号を返信する。
ファイルアクセス終了メッセージを受信するとファイルのロックを解除する。
3008
終了

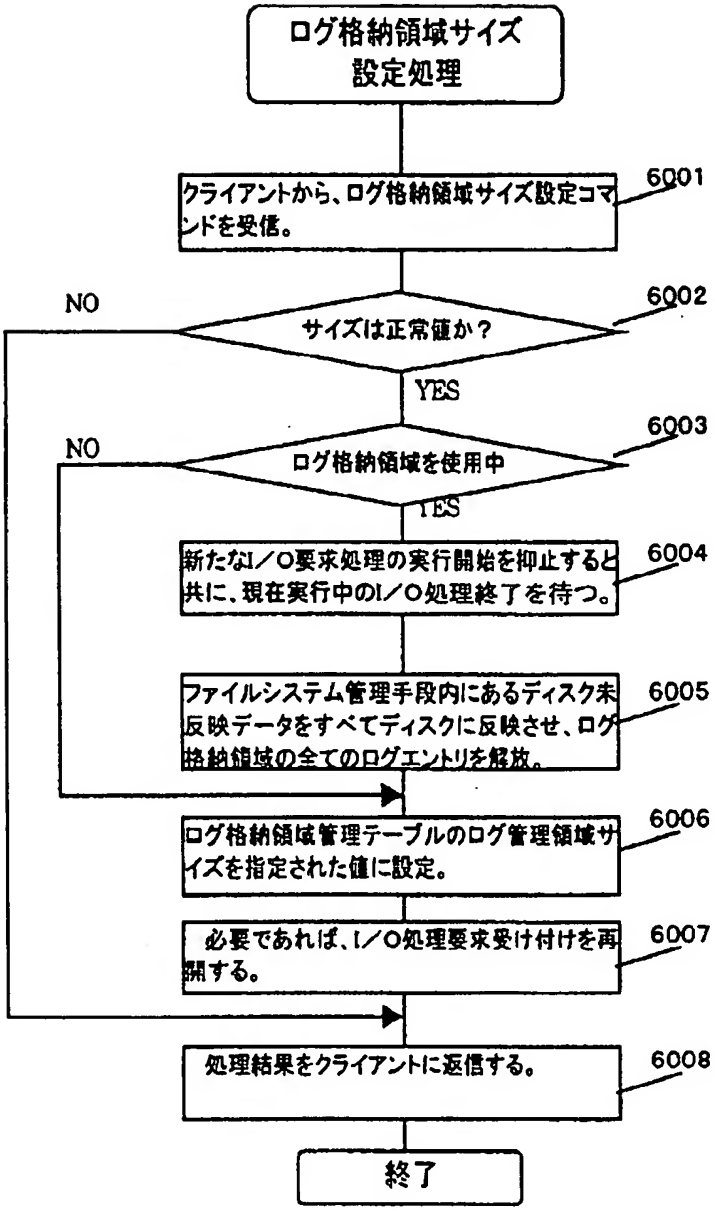
【図4】

図4


ファイルシステム復旧処理
ディスク番号を用いてディスク管理テーブルを検索し、直前の管理ファイルサーバ番号とログ種別を特定する。
4001
ログ種別がログ不採取？
4002
YES
NO
4010
fsck実行
管理ファイルサーバ番号を用いてログ格納領域テーブルを検索し、ログ格納領域アドレスとログ格納領域サイズを求める。
4003
ログ格納領域をスキャンして、対象ディスクに未反映の最も古いログエントリにポインタを設定。
4004
全てのログエントリの処理終了
4005
YES
NO
ログステータス情報が"データログ書き込み"
4006
NO
YES
ログエントリ内のファイルデータをディスクに書き込む。
4007
ログエントリ内の変更メタデータをファイルシステム管理情報に反映。
4008
ポインタを対象ディスクに未反映の次のログエントリに進める。
4009
終了

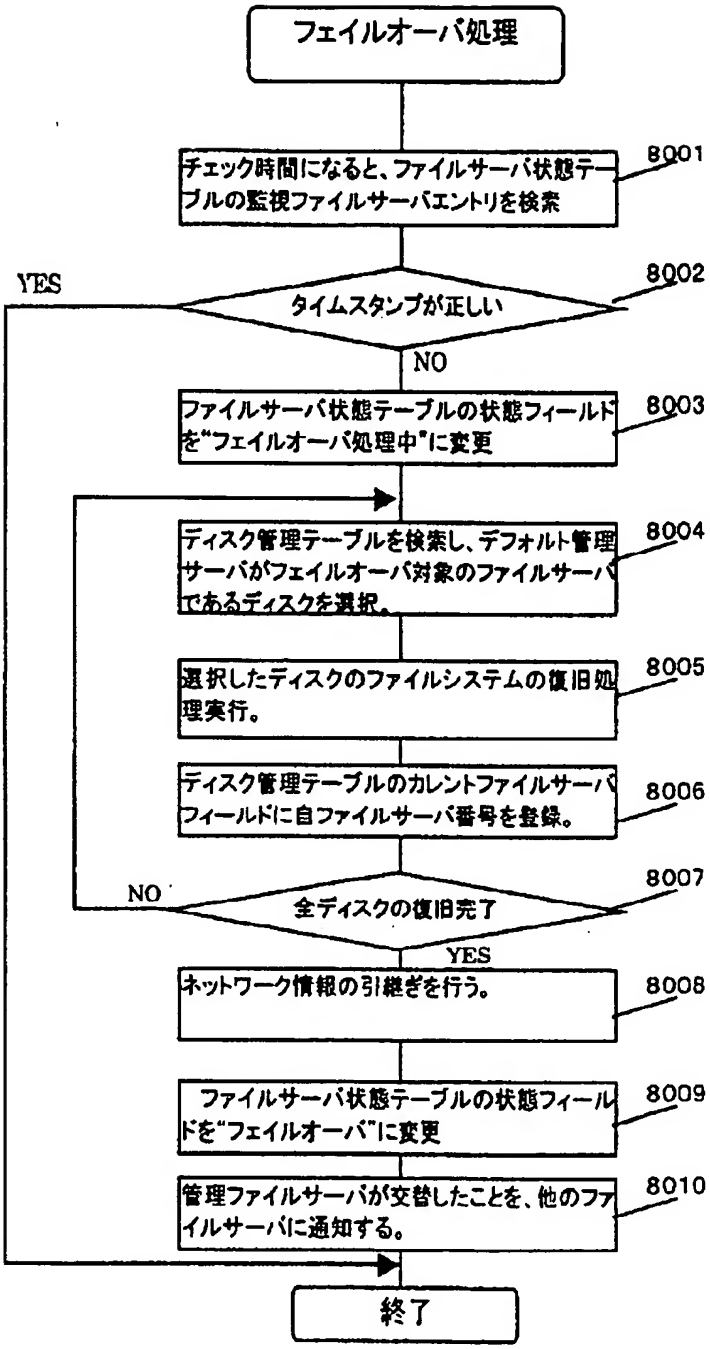
【図6】

図6


ログ格納領域サイズ
設定処理
クライアントから、ログ格納領域サイズ設定コマンドを受信。
6001
NO
サイズは正常値か？
6002
YES
NO
ログ格納領域を使用中
6003
YES
新たなI／O要求処理の実行開始を抑止すると共に、現在実行中のI／O処理終了を待つ。
6004
ファイルシステム管理手段内にあるディスク未反映データをすべてディスクに反映させ、ログ格納領域の全てのログエントリを解放。
6005
ログ格納領域管理テーブルのログ管理領域サイズを指定された値に設定。
6006
必要であれば、I／O処理要求受け付けを再開する。
6007
処理結果をクライアントに返信する。
6008
終了

【図８】

図8


フェイルオーバ処理
チェック時間になると、ファイルサーバ状態テーブルの監視ファイルサーバエントリを検索
8001
タイムスタンプが正しい
8002
YES
NO
ファイルサーバ状態テーブルの状態フィールドを"フェイルオーバ処理中"に変更
8003
ディスク管理テーブルを検索し、デフォルト管理サーバがフェイルオーバ対象のファイルサーバであるディスクを選択。
8004
選択したディスクのファイルシステムの復旧処理実行。
8005
ディスク管理テーブルのカレントファイルサーバフィールドに自ファイルサーバ番号を登録。
8006
全ディスクの復旧完了
8007
NO
YES
ネットワーク情報の引継ぎを行う。
8008
ファイルサーバ状態テーブルの状態フィールドを"フェイルオーバ"に変更
8009
管理ファイルサーバが交替したことを、他のファイルサーバに通知する。
8010
終了

フロントページの続き

(51)Int.Cl.[7]	識別記号	FI	テーマコード(参考)
G06F 3/06	540	G06F 3/06	540

(72)発明者 北村 学
神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72)発明者 高田 豊
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内

Fターム(参考) 5B065 BA01 CA18 ZA01 ZA15 ZA17
5B082 DD08 FA17 GB02 GB06 HA08